arXiv:1506.05506v2 [stat.ME] 7 Aug 2015

Noise Addition for Individual Records to Preserve Privacy and Statistical Characteristics: Case Study of Real Estate Transaction Data

Yuzo Maruyama Ryoko Tone and Yasushi Asami

The University of Tokyo
e-mail:
maruyama@csis.u-tokyo.ac.jp; ryo-t@ua.t.u-tokyo.ac.jp; asami@csis.u-tokyo.ac.jp

Abstract: We propose a new method of perturbing a major variable by adding noise such that results of regression analysis are unaffected. The extent of the perturbation can be controlled using a single parameter, which eases an actual perturbation application. On the basis of results of a numerical experiment, we recommend an appropriate value of the parameter that can achieve both sufficient perturbation to mask original values and sufficient coherence between perturbed and original data.

1. Introduction

Increasing amounts of information are now circulated because of recent advancements in digitalization, thereby increasing the importance of protecting personal information. Information that can identify a person should not be publicized or utilized without the person's consent. In the case of information regarding real estate, the location can be identified by combining several information sources, which in turn might be used to identify a person, such as an owner or resident. Spatial information of the sort that relates to real estate can be considered to require special protection.

Two factors are important in dealing with privacy-sensitive information. First, if information is leaked, then the organization responsible for the information risks receiving compensation claims because of privacy protection failure. Second, to avoid possible troubles due to potential information leaks, publicized data tend to become very rough or vague to avoid potential trouble, often hindering the usefulness of real estate analyses aimed at understanding the market.

A promising way of dealing with this situation is to protect personal information by adding noise to acquired data. A typical example of sensitive information is transaction data, which can include transacted prices, real estate or transacting person characteristics, and information regarding transaction conditions. Publicized data tend to omit information about characteristics of transacting persons, and hence, such contents are assumed not to be included in the database. In this case, one of the most sensitive types of data will be the transacted price. Private information will be protected if noise is added to the price data. However, tactless noise addition seriously distorts data analysis results. Therefore, providing a method of adding noise without distorting data analyses but still protecting privacy is very important. This study is devoted to proposing and applying such a method, assuming that the main concern of the analyses is hedonic analysis, i.e., regression analysis with the transacted price being the response variable.

Takemura (2003) reviewed statistical issues in publicizing individual data. He listed several methods of protecting personal information, such as (1) direct hiding by making the information secret, (2) global categorization by organizing values into several coarse classes, and (3) disturbance by replacing actual values with different ones (such as swapping by exchanging individual values, the post-randomization method [PRAM], or the addition of noise). Direct hiding and global categorization are not appropriate for releasing data for detailed analyses because the resolution of the information can become very coarse. Disturbance methods are superior in this aspect, although they usually introduce errors into analyses, and such effects must be carefully examined.

One well-known method of protecting personal information is the statistical disclosure limitation (SDL) method. SDL is a general term for methods of protecting identification of personal data by adding perturbations, modifications, or summarization (Shlomo (2010)). The main concern is to reduce identification risk as well as to retain data usability.

Typically, three kinds of methods are often used to reduce identification risk, including (1) methods of establishing coarse categorization, (2) methods of generating new data with statistical characteristics similar to those of the original data, and (3) methods of adding noise to the original data (Karr et al. (2006); Oganian and Karr (2011)).

Substantial research has been conducted on methods of establishing coarse categorization. In particular, population uniqueness, the feature that a combination of attributes becomes unique in the parent population, has been studied extensively. For example, Manrique-Vallier and Reiter (2012) estimated the risk of population uniqueness for discrete data.

Regarding methods of generating new data, the swapping method, in which categorical data are probabilistically exchanged, is well known. One such swapping method is PRAM, which perturbs the exchanging of categorical data (Gouweleeuw et al. (1998); Willenborg and Waal (2001)). In this method, a transition probability matrix is constructed and then used as the basis for exchanging categorical data, while maintaining the original proportions of the categories.

A variety of methods of adding noise, while carefully maintaining qualitative features, have been proposed. For example, Oganian and Karr (2011) focused on features such as the positivities of values and the magnitude relations between pairs of values. They proposed a method of adding noise such that the positivities of values, mean values, and variance-covariance matrices remain the same. One remarkable idea is to use multiplicative noise addition to avoid obtaining negative values. Moreover, they demonstrated the stability of results after regression analyses. A similar method of maintaining the characteristics of attributes was proposed by Abowd and Woodcock (2001). Another method of adding noise to avoid the risk of identification is to introduce random noise distributed following a peculiar symmetric distribution with a hole in the center. With this method, the perturbed value is never close to the original value, and therefore, the risk of identification is drastically reduced. In the actual application of this method, the noise distribution is not publicized, hindering analyses using the distribution (Reiter (2012)).

In general, noise addition can influence the quality of subsequent analyses. Fuller (1993) noted that noise addition has an influence similar to that of introducing measurement errors to explanatory variables. Several methods have been devised to minimize the influence of noise in particular analyses. For example, some methods maintain the original mean values and variance-covariance matrices (Ting, Fienberg and Trottini (2008); Shlomo and De Waal (2008)). In our paper, which focuses on regression analysis, a method is proposed in which adding noise produces robust results.

The paper is organized as follows. In Section 2, we propose a method of adding noise to a response variable and show that some important statistics do not change with noise addition. In Section 3, numerical experiments are conducted to examine how the results of multivariate analyses, apart from the assumed regression analysis, can change. Finally, Section 4 concludes with a summary and suggests possible extensions of our method.

2. Theoretical results

We assume that the $n \times (p+1)$ design matrix $\boldsymbol{X}$ is given by $(\mathbf{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$, where $\mathbf{1}_n$ is an n-dimensional vector of ones, and the n-dimensional response vector is $\boldsymbol{y}$. We also assume that n is sufficiently larger than p and that the rank of $\boldsymbol{X}$ is $p+1$. Then the ordinary least squares (OLS) estimator is

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)' = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}.$$

A decomposition of $\boldsymbol{y}$ based on the OLS estimator $\hat{\boldsymbol{\beta}}$ is $\boldsymbol{y} = \hat{\boldsymbol{y}} + \boldsymbol{e}$ where

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}'\boldsymbol{y}$$

is the predictive vector and

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I}_n - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1} \boldsymbol{X}')\boldsymbol{y}$$

is the residual vector. Then the coefficient of determination defined by

$$R^2 = 1 - \frac{\|\boldsymbol{e}\|^2}{\|\boldsymbol{y} - \bar{y}\mathbf{1}_n\|^2} \tag{2.1}$$

where $\bar{y}$ is the sample mean of $\boldsymbol{y}$, measures the goodness of fit resulting from the use of the OLS estimator $\hat{\boldsymbol{\beta}}$. The coefficient of determination, R^2, is hence regarded as a key quantity in regression analysis. The t-value of the regression coefficient β_j for $j = 0, 1, \ldots, p$, is another key quantity and is defined by

$$t_j = \frac{\sqrt{n-p-1}}{d_j} \frac{\hat{\beta}_j}{\|\boldsymbol{e}\|} \tag{2.2}$$

where d_j is the $(j+1)$-th diagonal component of $(\boldsymbol{X}'\boldsymbol{X})^{-1}$. When Gaussian linear regression is performed, t_j has a Student's t-distribution with $n-p-1$ degrees of freedom under the null hypothesis $\beta_j = 0$.

The objective of the derivation presented herein is to add perturbation to the original response vector and achieve tractable tuning of the R^2 and t-values. Any n-dimensional random vector

$$\boldsymbol{v} = (v_1, \ldots, v_n)'.$$

may be used as the starting point. Since n is sufficiently greater than p, $\boldsymbol{v}$ cannot be expressed as a linear combination of $\boldsymbol{e}, \boldsymbol{1}_n, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_p$ with probability one. In other words,

$$\boldsymbol{u} = \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \boldsymbol{e}\boldsymbol{e}'/\|\boldsymbol{e}\|^2\right)\boldsymbol{v}, \tag{2.3}$$

cannot be the zero vector. The noise vector considered in this paper is a linear combination of $\boldsymbol{e}$ and $\boldsymbol{u}$, given by

$$\boldsymbol{\epsilon} = \frac{a\|\boldsymbol{e}\|}{1+b}\left\{\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} + \sqrt{b}\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\}, \tag{2.4}$$

where $a \neq 0$ and $b \geq 0$. When $\boldsymbol{y} + \boldsymbol{\epsilon}$ is used instead of the original response vector $\boldsymbol{y}$, we have the following result.

Theorem 2.1. *1. The sample mean of $\boldsymbol{y} + \boldsymbol{\epsilon}$ is $\bar{y}$ for any a and b.*

2. *The OLS estimator for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ remains the same for any a and b, that is,*

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} + \boldsymbol{\epsilon}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

3. *The t-values for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ are given by*

$$\tilde{t}_j = \left\{\frac{1+b}{1+b+a(a+2)}\right\}^{1/2} t_j,$$

for $j = 0, \ldots, p$.

4. *The coefficient of determination for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ is*

$$\tilde{R}^2 = \left\{\frac{1+b}{1+b+a(a+2)(1-R^2)}\right\} R^2.$$

5. *The correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y} + \boldsymbol{\epsilon}$ is*

$$r_{y,y+\epsilon} = \frac{1+b+a(1-R^2)}{(1+b)^{1/2}\{1+b+a(a+2)(1-R^2)\}^{1/2}}.$$

Proof. By Part 3 of Lemma 2.1, we have $\boldsymbol{X}'\boldsymbol{\epsilon} = \boldsymbol{0}$, the first component of which is $\boldsymbol{1}_n'\boldsymbol{\epsilon} = 0$. Hence Part 1 follows.

Since $\boldsymbol{X}'\boldsymbol{\epsilon} = \boldsymbol{0}$, we have

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y}+\boldsymbol{\epsilon}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\epsilon} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \quad (2.5)$$

which completes the proof of Part 2.

Note that the t-values are defined by (2.2). By (2.5), any component of the OLS estimator keeps the same. Further $\sqrt{n-p-1}/d_j$ does not depend on the response vector. Hence Part 3 follows from Part 6 of Lemma 2.1.

Note the coefficient of determination is defined by (2.1). Since the sample mean of $\boldsymbol{y}+\boldsymbol{\epsilon}$ is also $\bar{y}$ as in Part 1 of this theorem, the coefficient of determination for the response vector $\boldsymbol{y}+\boldsymbol{\epsilon}$ is

$$1 - \frac{\|(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y}+\boldsymbol{\epsilon})\|^2}{\|\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n\|^2},$$

which is rewritten as

$$1 - \frac{\{1 + a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + \{a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2}$$

by Parts 5 and 6 of Lemma 2.1. By the definition of R^2, we have

$$1 - R^2 = \|\boldsymbol{e}\|^2/\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2, \quad (2.6)$$

which completes the proof of Part 4.

The correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y}+\boldsymbol{\epsilon}$ is

$$\frac{(\boldsymbol{y}-\bar{y}\boldsymbol{1}_n)'(\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n)}{\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|\|\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n\|}.$$

By Parts 3 and 4 of Lemma 2.1 as well as (2.6), we have

$$\begin{aligned}(\boldsymbol{y}-\bar{y}\boldsymbol{1}_n)'(\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\boldsymbol{1}_n) &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + (\boldsymbol{y}-\bar{y}\boldsymbol{1}_n)'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + (\hat{\boldsymbol{y}}+\boldsymbol{e}-\bar{y}\boldsymbol{1}_n)'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + (\boldsymbol{X}\hat{\boldsymbol{\beta}}-\bar{y}\boldsymbol{1}_n)'\boldsymbol{\epsilon} + \boldsymbol{e}'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2 + \boldsymbol{e}'\boldsymbol{\epsilon} \\ &= \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2\left[1 + \{a/(1+b)\}(1-R^2)\right].\end{aligned}$$

Further, by Part 5 of Lemma 2.1, we have

$$\|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n+\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y}-\bar{y}\boldsymbol{1}_n\|^2\left[1 + \{a(a+2)/(1+b)\}(1-R^2)\right],$$

which completes the proof of Part 5. □

The lemma below summarizes fundamental properties related to $\boldsymbol{e}$ and $\boldsymbol{\epsilon}$, which are needed in the proof of Theorem 2.1.

Lemma 2.1. *1. $\boldsymbol{e}$ is orthogonal to $\mathbf{1}_n, \boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ or equivalently $\boldsymbol{X}'\boldsymbol{e} = \mathbf{0}$.*

2. $\boldsymbol{u}$ is orthogonal to $\boldsymbol{e}, \mathbf{1}_n, \boldsymbol{x}_1, \dots, \boldsymbol{x}_p$ or equivalently $\boldsymbol{X}'\boldsymbol{u} = \mathbf{0}$ and $\boldsymbol{e}'\boldsymbol{u} = 0$.

3. $\boldsymbol{X}'\boldsymbol{\epsilon} = \mathbf{0}$.

4. $\boldsymbol{e}'\boldsymbol{\epsilon} = a\|\boldsymbol{e}\|^2/(1+b)$ and $\|\boldsymbol{\epsilon}\|^2 = a^2\|\boldsymbol{e}\|^2/(1+b)$.

5. The sum of squared deviation of $\boldsymbol{y}+\boldsymbol{\epsilon}$ is

$$\|\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\mathbf{1}_n\|^2 = \|\boldsymbol{y}-\bar{y}\mathbf{1}_n\|^2 + \{a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2.$$

6. The residual sum of squares for $\boldsymbol{y}+\boldsymbol{\epsilon}$ is

$$\|(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y}+\boldsymbol{\epsilon})\|^2 = \{1 + a(a+2)/(1+b)\}\,\|\boldsymbol{e}\|^2.$$

Proof. Since $\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{X}'$, we have

$$\begin{aligned}\left(\mathbf{1}_n'\boldsymbol{e}\ \boldsymbol{x}_1'\boldsymbol{e}\ \cdots\ \boldsymbol{x}_p'\boldsymbol{e}\right)' &= \boldsymbol{X}'\boldsymbol{e} \\ &= \boldsymbol{X}'\left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y} \\ &= \left(\boldsymbol{X}' - \boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\right)\boldsymbol{y} \\ &= \mathbf{0},\end{aligned}$$

which completes the proof of Part 1. In the same way, Part 2 can be proved.

Recall $\boldsymbol{\epsilon}$ is given by a linear combination of $\boldsymbol{e}$ and $\boldsymbol{u}$,

$$\boldsymbol{\epsilon} = \frac{a\|\boldsymbol{e}\|}{1+b}\left\{\frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} + \sqrt{b}\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|}\right\}. \tag{2.7}$$

Then Part 3 follows from Parts 1 and 2. Part 4 follows from the orthogonality of $\boldsymbol{e}$ and $\boldsymbol{u}$ together with (2.7).

Since the sample mean of $\boldsymbol{y}+\boldsymbol{\epsilon}$ is $\bar{y}$ by Part 1 of Theorem 2.1, the sum of squared deviation of $\boldsymbol{y}+\boldsymbol{\epsilon}$, $\|\boldsymbol{y}+\boldsymbol{\epsilon}-\bar{y}\mathbf{1}_n\|^2$, is expanded as

$$\|\boldsymbol{y}-\bar{y}\mathbf{1}_n\|^2 + 2(\boldsymbol{y}-\bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} + \|\boldsymbol{\epsilon}\|^2.$$

By Part 3, we have

$$(\boldsymbol{y}-\bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} = (\hat{\boldsymbol{y}}+\boldsymbol{e}-\bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} = (\boldsymbol{X}\hat{\boldsymbol{\beta}}+\boldsymbol{e}-\bar{y}\mathbf{1}_n)'\boldsymbol{\epsilon} = \boldsymbol{e}'\boldsymbol{\epsilon} = a\|\boldsymbol{e}\|^2/(1+b).$$

Then Part 5 follows from Part 4.

Since $\boldsymbol{X}'\boldsymbol{\epsilon} = \mathbf{0}$ by Part 3, we have

$$(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{y}+\boldsymbol{\epsilon}) = \boldsymbol{e}+\boldsymbol{\epsilon}.$$

From Part 4, the residual sum of squares is

$$\|\boldsymbol{e}+\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{e}\|^2 + 2\boldsymbol{e}'\boldsymbol{\epsilon} + \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{e}\|^2 + 2a\|\boldsymbol{e}\|^2/(1+b) + a^2\|\boldsymbol{e}\|^2/(1+b),$$

which completes the proof of Part 6. □

By Theorem 2.1, we see that $a = -2$ is a special case, as follows.

Theorem 2.2. *Assume $a = -2$. Then, we have the followings.*

1. *For any $b > 0$, the coefficient of determination for $\boldsymbol{y} + \boldsymbol{\epsilon}$ is equal to R^2, the coefficient of determination for the original $\boldsymbol{y}$.*
2. *For any $b > 0$, the t-value of β_j $(j = 0, 1, \dots, p)$ for the response vector $\boldsymbol{y} + \boldsymbol{\epsilon}$ is equal to t_j.*
3. *The correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y} + \boldsymbol{\epsilon}$ is*

$$r_{y,y+\epsilon} = 1 - \frac{2(1 - R^2)}{1 + b}. \tag{2.8}$$

Recall that $\boldsymbol{\epsilon}$ is a function of $\boldsymbol{v}$, any random n-dimensional vector, through the relationships, (2.3) and (2.4), that is,

$$\boldsymbol{u} = \left(\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' - \frac{\boldsymbol{e}\boldsymbol{e}'}{\|\boldsymbol{e}\|^2} \right) \boldsymbol{v}, \ \boldsymbol{\epsilon} = \frac{a\|\boldsymbol{e}\|}{1 + b} \left\{ \frac{\boldsymbol{e}}{\|\boldsymbol{e}\|} + \sqrt{b} \frac{\boldsymbol{u}}{\|\boldsymbol{u}\|} \right\}.$$

In Parts 1 and 2 of Theorem 2.2, the choice $a = -2$ guarantees that the coefficient of determination and t-value remain the same regardless of $\boldsymbol{v}$.

By Part 3 of Theorem 2.2, $r_{y,y+\epsilon}$ increases with b for fixed R^2. The correlation coefficients between the original responses $\boldsymbol{y}$ and perturbed responses $\boldsymbol{y} + \boldsymbol{\epsilon}$ with $a = -2$, varying $b \geq 0$ and R^2, are illustrated in Table 1.

In actual application, it is desirable to have relatively high correlation, because data users might assume that the perturbed response is close to the original response. However, if the correlation is very high, then the perturbed response is very close to the original response, and the objective of concealing the actual response cannot be achieved. Thus, it is necessary to determine a value of b that prevents the perturbed response from being too close to the actual response, as will be discussed through the analysis of real data in the next section.

Remark 2.1. When $a = -2$ and $b = 0$, we have $\boldsymbol{\epsilon} = -2\boldsymbol{e}$ as the noise or, equivalently

$$\boldsymbol{y} - 2\boldsymbol{e} = \hat{\boldsymbol{y}} - \boldsymbol{e} \tag{2.9}$$

as the perturbed response. In this case, it is clear that the coefficient of determination and t-value remain the same, since y_i and $y_i - 2e_i$ for $i = 1, \dots, n$ are symmetric with respect to the point $\hat{y}_i = y_i - e_i$. Since the noise $\boldsymbol{\epsilon} = -2\boldsymbol{e}$ does not depend on $\boldsymbol{v}$, there is no randomness in the noise. Theorem 2.2 ensures that, for random $\boldsymbol{v}$, as in (2.4), it is possible to construct the noise $\boldsymbol{\epsilon}$ such that the coefficient of determination and t-value remain the same.

Remark 2.2. As in Theorem 2.2, the choice $a = -2$ with random $\boldsymbol{v}$ was surprisingly found to retain the R^2 and t values. Following are some remarks for the other choices. For $a \in (-\infty, -2) \cup (0, \infty)$, both R^2 and the absolute value of t values are reduced. For example, $b > 0$ and $a = -1 \pm \sqrt{b + 2} \in (-\infty, -2) \cup (0, \infty)$ yield

$$\tilde{t}_j = \frac{1}{\sqrt{2}} t_j, \tilde{R}^2 = \frac{R^2}{2 - R^2} < R^2. \tag{2.10}$$

Table 1
Correlation coefficient of $\boldsymbol{y}$ and $\boldsymbol{y} + \boldsymbol{\epsilon}$ with $a = -2$

$R^2 \backslash b$	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0
0.4	-0.2	0.04	0.2	0.31	0.4	0.47	0.52	0.56	0.6
0.6	0.2	0.36	0.47	0.54	0.6	0.64	0.68	0.71	0.73
0.8	0.6	0.68	0.73	0.77	0.8	0.82	0.84	0.85	0.87

Note that R^2 and t values can be completely controlled. Thus the data provider safely provide data with the relation between $\{t_j, R^2\}$ and $\{\tilde{t}_j, \tilde{R}^2\}$ described by (2.10), and practitioners can restore the original R^2 and t-value independently. An efficient method of opening data with reduced accuracy will be reported elsewhere.

3. Numerical experiment

In the previous section, a method was proposed to add noise to the response variable. This method can be applied when a real estate database is released into the public domain, by adding noise to the transacted price, which is considered to be sensitive information in Japan. As Theorem 2.2 in the previous section ensures, the results of regression analysis using the perturbed data will not change. However, in actual application, a variety of analyses will be devised, and the theorems do not apply in cases with unexpected applications. Thus, it is necessary to verify whether the proposed method remains appropriate even in unexpected applications.

The precision of the results might be degraded if analytical operations not assumed in the theory are applied. In such cases, permissible error levels resulting from the perturbation must be determined. In the following, a numerical experiment to determine the relationship between perturbation and precision level is discussed.

3.1. Data used in the experiment

The data source used for the numerical experiment was At Home Co. Ltd. The data contained real estate advertisement information from 2008. The database for the experiment was created by supplementing some spatial variables. It contained $1,320$ cases of newly built detached houses in Setagaya Ward in Tokyo Prefecture[1]. The variables included the price of the property (yen), the time to the nearest railway station (minutes), a dummy variable representing bus usage, the area of the site (square meters), the floor area (square meters), a dummy variable signifying leased land, the designated building coverage ratio,

[1] We selected data that contained information about the designated floor area and building coverage ratios. Such data are thought to be important in real estate analysis in Japan. In the original database, a new record was added each time a property owner changed the price in the advertisement. In such situations, we selected only the newest record.

the designated floor area ratio, the time to Shinjuku by rail from the nearest station (minutes), the time to Shibuya by rail (minutes), the time to Yokohama by rail (minutes), the time to Tokyo by rail (minutes), the width of the nearest road (meters), and a dummy variable signifying the nearest road to the south of lot. Note that Shinjuku, Shibuya, Yokohama, and Tokyo are four major railway stations in the study region. Among these variables, the times to the railway stations, width of the nearest road, and dummy variable signifying the nearest road to the south are spatial variables, as described in the next subsection.

3.2. Creation of spatial variables

The times to the major railway stations from the nearest station; the width of the nearest road, as measured from the representative point of the property; and the dummy variable signifying whether the nearest road is located to the south of the property were added to the original database as variables for signifying spatial relationships. The width of the nearest road to the representative point of the property was regarded as the width of the nearest road, which was done because precise digital data for lots are not available. Accordingly, the dummy variable signifying whether the nearest road was located to the south of the property was regarded as the dummy variable signifying the nearest road to the south of lot.

The times to the major railway stations from the nearest station were calculated using the search system for guiding transferring railways provided by NAVITIME Japan Co. Ltd. This system automatically calculates the time required to travel to the major railway stations, i.e., Shinjuku, Shibuya, Yokohama, and Tokyo, from the railway station nearest to the property. To determine the times required in this study, the departing time was set to $12:00$ (noon) on August 2, 2010.

The width of the nearest road from the representative point of the property was calculated as follows. Mapple 10000 digital data produced by Shobunsha Publications Inc. contain digital road data classified by road width categories, such as 4-5m and 5-6m. The median of each class was assigned as the road width. For example, a width of 4.5m was used for the 4-5m class. With the geographic information system (GIS) software ArcGIS 10, the nearest road was assigned for each property, and the width of the road calculated as described above was set to be the width of road nearest to the representative point of the property.

In the real estate market in Japan, a residential lot tends to be evaluated highly if it is adjacent to a road to the south of lot, because receiving substantial sunlight is preferred in Tokyo. For example, The Real Estate Transaction Modernization Center (1986) treats properties adjacent to roads to the south of lots more favorably in their property appraisals. With this preference in mind, the dummy variable signifying whether the nearest road is located to the south of the property was also added to the database.

This dummy variable was constructed as follows. Using ArcGIS 10, the direction to the nearest road was calculated, such that 0° was located to the east, and

Table 2
Summary of variable statistics

	min	max	mean	s.d.
price of the property (yen)	34800000	330000000	72431491	25539447
time to the nearest railway station (minutes)	0	25	10.60	4.83
d.v.[a] representing bus usage	0	1	0.07	0.26
area of the site (square meters)	29.53	211.49	88.56	25.48
floor area (square meters)	47.07	228.48	98.94	20.06
d.v.[a] signifying leased land	0	1	0.03	0.17
designated building coverage ratio	40	80	54.18	7.70
designated floor area ratio	80	300	141.43	47.10
time to Shinjuku by rail (minutes)	5	32	18.72	5.29
time to Shibuya by rail (minutes)	3	29	14.86	6.01
time to Yokohama by rail (minutes)	17	64	44.30	10.99
time to Tokyo by rail (minutes)	23	48	34.09	4.90
width of the nearest road (meters)	4.5	35	5.80	2.25
d.v.[a] signifying the nearest road to the south of lot	0	1	0.28	0.45

[a] d.v. stands for “dummy variable”.

the value increased to 180° counterclockwise and decreased to $-180°$ clockwise. The range from $-135°$ to $-45°$ was judged to be to the south, in which case the dummy variable was set to one, and it was set to zero otherwise. The statistics of the variables are summarized in Table 2.

3.3. Numerical experiment with perturbed property price

The perturbed property price, which was generated by adding noise to the response variable using the method described in the previous section, was numerically tested as described in this subsection. The explanatory variables used were the 13 variables in Table 2.

3.3.1. Statistics of the perturbed property price

Although Part 1 of Theorem 2.1 guarantees that the mean of the perturbed response variable is exactly equal to the mean of the original response, the equality or similarity of the other statistics, such as the minimum value, maximum value, and first and third quantiles, theoretically cannot be controlled. In this section, the generation of four sets of quasi-response variables with different $\boldsymbol{v}$ values, $a = -2$, and $b = 1$ is described, to analyze the degrees of perturbation of the statistics among the five sets, including the original response (original, quasi1, quasi2, quasi3, and quasi4).

Figure 1 shows boxplots of the five sets. When the original and quasi-response variables are compared, the medians are very similar, but the quantiles, minima, and maxima are quite different. It is also evident that, among the four sets of quasi variables, all of the statistics are similar. Figure 2 shows scatterplots of the original variables and of the four sets of quasi variables. Although the plots for the four sets of quasi variables appear very similar, the different $\boldsymbol{v}$ values

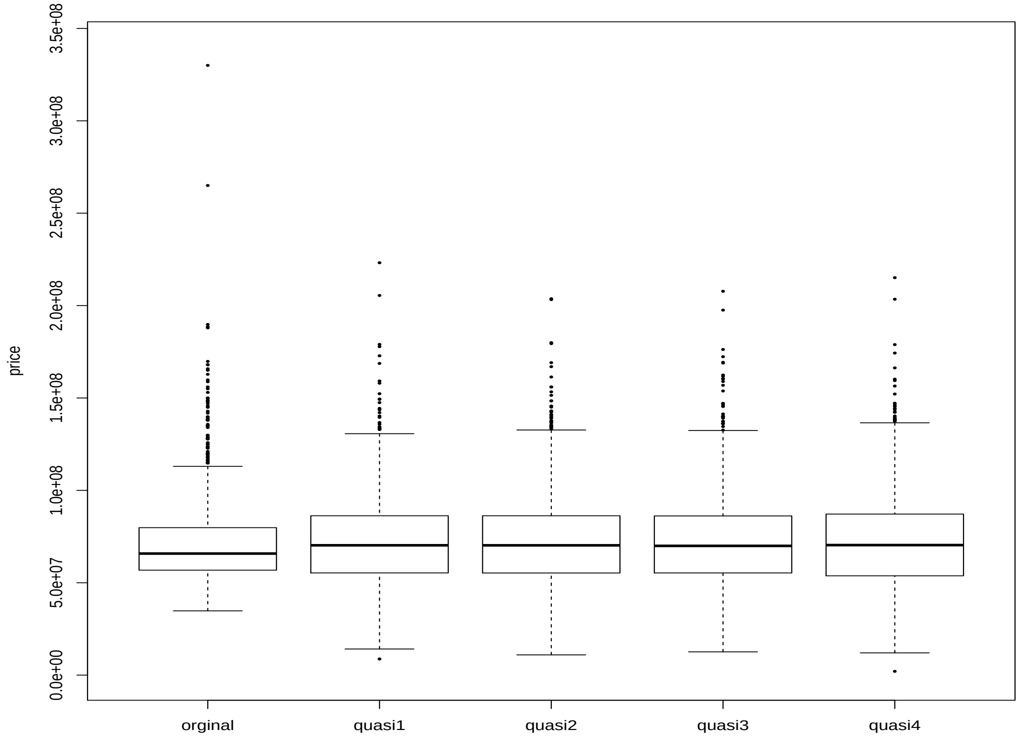

Fig 1. *Boxplots of five sets of response variables*

imply different quasi variables, as explained in Section 2. Table 3 provides the correlation matrix for the five sets of response variables. By Part 3 of Theorem 2.2, the correlation coefficient between the original and quasi variables is given theoretically by $1 - 2(1 - R^2)/(1 + b)$, which equals R^2 for $b = 1$. Among the quasi variables, the correlations in all cases are approximately 0.78.

Remark 3.1. In this particular data set, the response variable was the property price, which was expected to be positive. Hence, a positive perturbed price is strongly desirable. As claimed in Remark 2.1, for sufficiently small b, we have

$$\boldsymbol{y} + \boldsymbol{\epsilon} \approx \hat{\boldsymbol{y}} - \boldsymbol{e}.$$

Suppose there exist individuals i with relatively expensive prices y_i, when relatively lower prices y_i are expected. Then e_i increases, and as a result

$$y_i + \epsilon_i \approx \hat{y}_i - e_i < 0 \tag{3.1}$$

can occur. In our data set, such situations rarely occurred for $b = 1.2$ or less and never occurred for $b = 1.3$ or greater. To the best of our knowledge, the

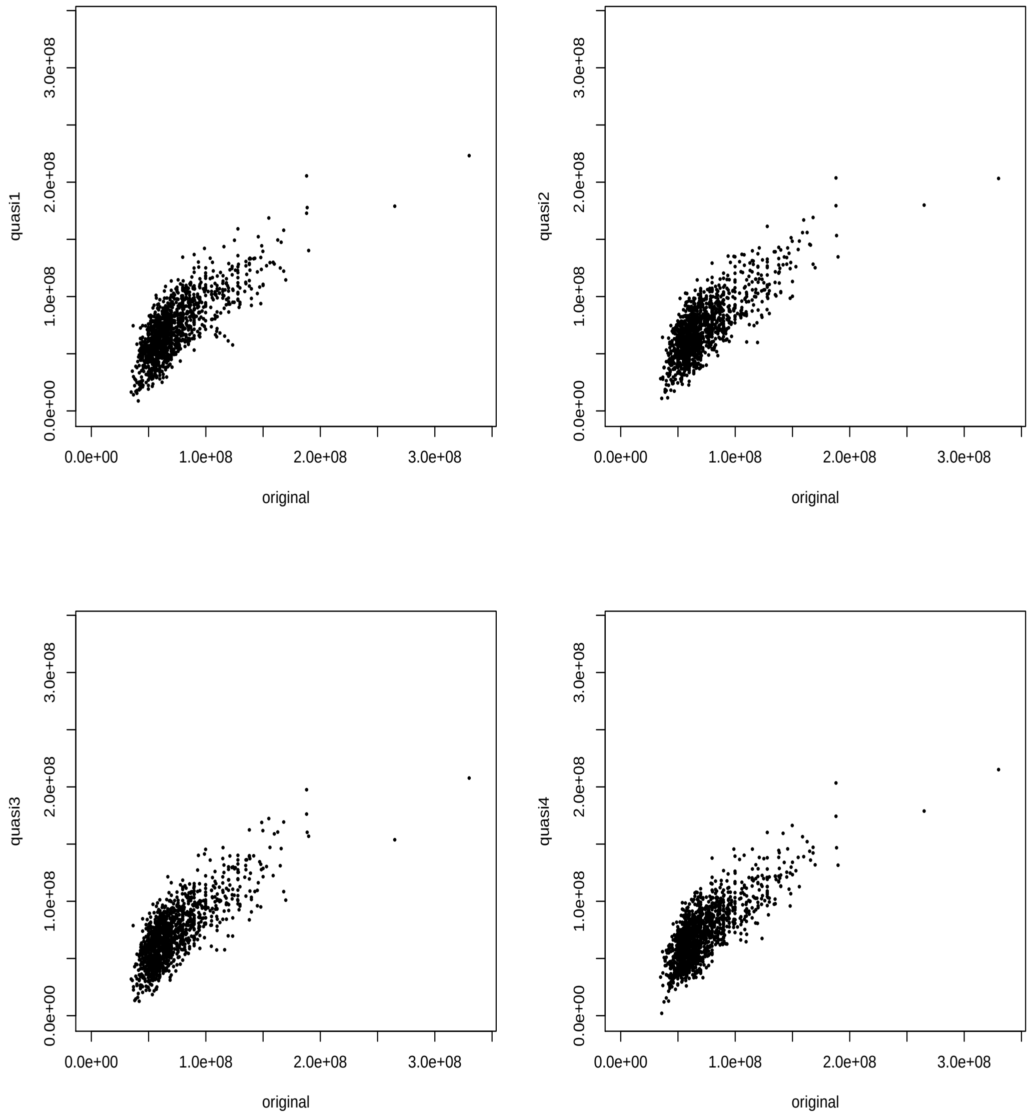

Fig 2. *Scatterplots of five sets of response variables*

Table 3
Correlations between original response and four sets of quasi responses

	orig	quasi1	quasi2	quasi3	quasi4
orig	1	0.7748	0.7748	0.7748	0.7748
quasi1	0.7748	1	0.7693	0.7749	0.7814
quasi2	0.7748	0.7693	1	0.7870	0.7812
quasi3	0.7748	0.7749	0.7870	1	0.7733
quasi4	0.7748	0.7814	0.7812	0.7733	1

occurrence of such situations is theoretically not controllable through the choices of b and $\boldsymbol{v}$. When (3.1) occurs, it is recommended to generate $\boldsymbol{\epsilon}$ with different $\boldsymbol{v}$ values until $\min\{y_i + \epsilon_i\} > 0$ is achieved.

3.3.2. Regression analysis using only a portion of the database

The theory assumes that all of the data will be used for the analysis. If only a portion of the perturbed data is used, then the theorems do not apply exactly. In actual analyses for real estate data, only a portion of the (perturbed) database is used for the analysis. In such cases, it is necessary to know how the results might differ from the theoretical results and to follow the subsequently described guidelines to choose an appropriate value of b.

For this purpose, a critical value of b may be obtained such that the difference between the regression model using the perturbed property price as the response variable and the original property price is not statistically significant.

3.3.3. Chow test

From $1,320$ cases (total database), 20% (i.e., 264 cases) were selected randomly, and perturbed prices were generated for 13 b values (i.e., $b = 0.5$, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 2.0, 2.5). The Chow test was applied to determine whether the regression models with the original and the perturbed prices could be regarded as the same model. For each value of b, $1,000$ independently chosen samples were created and analyzed. As a result, for each value of b, $1,000$ values of the Chow test F value were derived. Ordering these values by magnitude, 5%, 10%, 50%, 90% and 95% (i.e., the 50^{th}, 100^{th}, 500^{th}, 900^{th} and 950^{th} value) of the points of the F value were derived. Figure 3 shows that larger values of b correspond to smaller F-value variations. Given the objective of choosing an appropriate value of b to generate a properly perturbed property price, the minimum value of b for which the null hypothesis of the Chow test (namely, H_0: "There is no statistically significant difference between two models") is not rejected can be considered the critical value of b. Note that the F value in the F distributions with degrees of freedom 14 and 500 that achieves a significance level 0.05 is $F = 1.71$. Hence, if F is less than 1.71, then the null hypothesis cannot be rejected, and therefore the two models can be regarded as statistically the same.

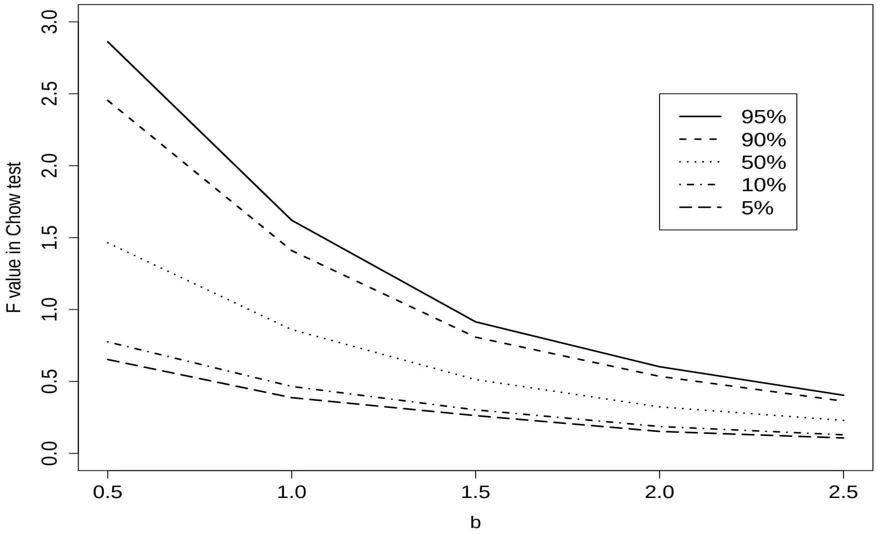

Fig 3. *The relation between F value and b*

For each value of b, the percentage of F values among $1,000$ trials that satisfied the acceptance condition of F less than 1.71 was calculated. In our numerical experiment, these percentages are 65.0%, 97.0%, and 100% for $b = 0.5$, $b = 1.0$, and $b \geq 1.4$, respectively, as seen in Table 4.

3.3.4. Recommended standard for b value

In the numerical experiment described above, when 20% was selected randomly, the Chow tests used to test the identity of the two models, that is, the regression models with the actual and perturbed property prices as the response variables, demonstrated that the F value satisfied the acceptance condition with 97.0% probability when $b = 1.0$ and 100% probability when $b \geq 1.4$. Assuming that approximately 5% is the permissible level for hypothesis rejection (i.e., that two models cannot be regarded as the same), $b = 1.0$ is judged appropriate, as it ensures that the perturbed price is perturbed sufficiently and, nonetheless, that the regression model with the perturbed price can be regarded as identical to the regression model with the original price. The appropriate value of b differs if another percentage is used to select the sample. For instance, we let q be the percentage used to select the sample and, using the above numerical experiment, we let $q = 0.2$ (20%). Assuming a 5% rejection level, the critical value of b, b_*, such that for b less than b_* the probability of rejection becomes greater than 5%, was calculated by changing q. Table 4 summarizes the results. For all q values investigated in this study, $b_* = 1.0$ appears to be a reasonable choice, as

Table 4
Percentages of samples that accepted the null hypothesis, for which the perturbed sample can be regarded as statistically identical to the original sample for sample selection percentage, q, and b value

$b\backslash q$	0.05	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90
0.5	0.591	0.576	0.650	0.728	0.827	0.918	0.969	0.992	0.996	1.000
0.6	0.686	0.670	0.744	0.808	0.903	0.947	0.988	0.997	0.999	1.000
0.7	0.729	0.764	0.808	0.873	0.943	0.976	0.994	1.000	0.999	1.000
0.8	0.815	0.845	0.858	0.928	0.966	0.988	0.996	1.000	1.000	1.000
0.9	0.867	0.867	0.931	0.966	0.989	1.000	0.998	1.000	1.000	1.000
1.0	0.930	0.925	0.970	0.987	0.995	0.998	1.000	1.000	1.000	1.000
1.1	0.960	0.968	0.983	0.998	0.997	1.000	1.000	1.000	1.000	1.000
1.2	0.978	0.987	0.994	0.999	0.999	1.000	1.000	1.000	1.000	1.000
1.3	0.994	0.994	0.995	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.4	0.996	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.5	0.999	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

it balances the similarity and the perturbation to the original price.

4. Conclusion

This paper proposed a new method of perturbing a major variable by adding noise, while ensuring that the results of regression analysis are not affected. The extent of the perturbation can be controlled using a single parameter, b, which eases actual perturbation application. Moreover, $b = 1.0$ can be regarded as an appropriate value for achieving both sufficient perturbation to mask the original values and sufficient coherence between the perturbed and original data.

The proposed method masks only one major variable, but in actual application, many situations may be encountered in which only one variable is critical to put the entire dataset in the public domain. Our method will be useful in such situations. There are other possible uses of perturbed data, and the appropriateness of the b value must be examined by testing a greater variety of data-use cases. Admittedly, application of the proposed method is limited, because other variables are assumed to retain their original values. Thus, further methods of perturbing the explanatory variables are necessary to broaden the range of applications. Such extensions will be provided in subsequent work.

Acknowledgements

This work was partly supported by a Research Grant from the Secom Science and Technology Foundation. The work of the first and third authors was partly supported by Grant-in-Aids for Scientific Research No. 25330035 and No. 26590036, respectively, from the Japan Society for the Promotion of Sciences.

References

Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* 215–277.

Fuller, W. A. (1993). Masking procedures for microdata disclosure limitation. *Journal of Official Statistics* **9** 383–383.

Gouweleeuw, J. M., Kooiman, P., Willenborg, L. C. R. J. and de Wolf, P. P. (1998). Post randomisation for statistical disclosure control: Theory and implementation. *Journal of Official Statistics* **14** 463–478.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P. and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* **60** 224–232.

Manrique-Vallier, D. and Reiter, J. P. (2012). Estimating identification disclosure risk using mixed membership models. *Journal of the American Statistical Association* **107** 1385–1394.

Oganian, A. and Karr, A. F. (2011). Masking methods that preserve positivity constraints in microdata. *Journal of Statistical Planning and Inference* **141** 31–41.

Reiter, J. P. (2012). Statistical approaches to protecting confidentiality for microdata and their effects on the quality of statistical inferences. *Public opinion quarterly* **76** 163–181.

Shlomo, N. (2010). Releasing microdata: Disclosure risk estimation, data masking and assessing utility. *Journal of Privacy and Confidentiality* **2** 73–91.

Shlomo, N. and De Waal, T. (2008). Protection of Micro-data Subject to Edit Constraints Against Statistical Disclosure. *Journal of Official Statistics* **24** 229–253.

Takemura, A. (2003). Current trends in theoretical research of statistical disclosure control problem. *Proceedings of the Institute of Statistical Mathematics* **51** 252.

The Real Estate Transaction Modernization Center (1986). How to evaluate the price for residential area: Manual for appraising land price Technical Report, The Real Estate Transaction Modernization Center, Tokyo. Real estate properties distribution series, No.8.

Ting, D., Fienberg, S. E. and Trottini, M. (2008). Random orthogonal matrix masking methodology for microdata release. *International Journal of Information and Computer Security* **2** 86–105.

Willenborg, L. and Waal, T. (2001). Elements of statistical disclosure control.